Conditional Recall

Christoph Schlegel & Xinyuan Sun

November, 2024

In the neon-lit nights of 2025, Johnson & Johnson unveiled X. A pill, not larger than a snowflake, that promised a tempest of change. This miraculous drug didn't just allow people to cherry-pick memories to erase from their minds, it could also leave a reminder of this erasure in the minds of those who ingested it.

Amidst the iconic red-bricked walls of Harvard Law, you, with books in one hand and dreams in the other, are on a mission. You are not just another student; you carry the hope of revolutionizing the archaic chambers of the legal world. Each night, as you pore over the tomes of law, you wonder what greatness society can achieve.

On a cold evening, your phone buzzes. It's Dex, your old college friend turned underground dealer. His message is simple: "Got X. Special price for you." The temptation swirls around you. Would you trade the lessons of the past for a clearer, yet incomplete future? The decision rests in your hands.

1 Introduction

Perfect recall is a foundational assumption in extensive form games since Kuhn (1953). While this assumption naturally models human cognition — as humans cannot selectively forget information — it may be unnecessarily restrictive for artificial agents. In particular, Large Language Models (LLMs) can implement conditional recall through context segmentation: multiple instances can each hold private information, verify each other's reasoning chains for information leakage, and commit to forgetting the interaction. Credibility of commitments to forget can be enforced, for example, through the use of Trusted Execution Environments (TEEs). These capabilities suggests new mechanism design possibilities.

We formalize this insight through a game-theoretic framework in which players can choose to forget information they have previously learned while maintaining time consistency. Using a hypothetical "pill X" as our primitive, we characterize several classes of games where conditional recall enables more efficient equilibria: (i) information markets where Arrow's paradox previously prevented trade, (ii) information sharing schemes where one-time use of information is enforced, (iii) bargaining protocols where reputation effects sustain inefficient outcomes, and (iv) commitment devices where time inconsistency leads to suboptimal choices.

Through a series of strategic scenarios — from corporate negotiations to military intelligence — we demonstrate how conditional recall mechanisms strictly dominate classical solutions that assume perfect recall. Our results suggest that as artificial agents increasingly participate in strategic interactions, mechanism designers should exploit their ability to selectively forget information, a capability that transcends human cognitive constraints.

2 Model and Framing

We consider extensive form games of incomplete information. Such game $\Gamma = (N, A, \mathcal{H}, \mathcal{I}, u)$ consists[1] of a finite set of players N (including a player called "nature"), a set of **actions** A, a set $\mathcal{H}$ of sequences of actions, called **histories**, such that if $a_1, \ldots, a_K$ is a history and $L < K$ then $a_1, \ldots, a_L$ is also a history,[2] payoff functions $u_i : Z \to \mathbb{R}$ for each $i \in N$ where Z is the set of terminal histories (those histories that cannot be further extended), a mapping from non-terminal histories to players and a mapping from non-terminal histories to actions which prescribe which player takes an action after the history and which actions are available to them (denote by $\mathcal{H}_i$ for each $i \in N$ the set of histories after which i takes an action and $A(h) \subseteq A$ to be the set of available actions after history h) and a partition of $\mathcal{H}_i$ for each i into **information sets** $\mathcal{H}_i = \bigcup_{I \in \mathcal{I}_i} I$ such that the set of available actions are the same after each history in the same information set. Information sets model which states of the world are indistinguishable for the players when choosing an action, based on the information that they take into consideration when taking an action. In the following, for $h, h' \in \mathcal{H}$ we write $h' \succ h$ if h' is an extension of h and for $I, J \in \mathcal{I}_i$ we write $J >_i I$ whenever there is a $h \in I$ and a $h' \in J$ with $h \succ h'$. Note that $\succ$ is a strict partial order, i.e. it is transitive and irreflexive, whereas $>_i$ is not necessarily a strict partial order.

What does it mean to forget in an extensive form game? For a given history h and player i we can consider the sequence $X_i(h)$ of information sets he has encountered and the actions he has chosen along the history h. We say that player i has **Perfect Recall** if $X_i(h) = X_i(h')$ whenever h and h' are in the same information set. Thus, if two histories are in the same information set for player i, he couldn't distinguish between them even if he takes into consideration his entire experience of playing the game up to that information set. Perfect recall is equivalent (see Ritzberger (1999)) to the simultaneous fulfillment of three independent properties: Players never forget what they did; they never forget what they knew;[3] and (for a given player i), past, present, and future have unambiguous meaning. The last requirement means that for i the relation $>_i$ is a strict partial order. If $>_i$ is transitive[4] but not irreflexive this is commonly called **absent-mindedness**, i.e. we are in a situation where i does not remember whether he already took an action or not. In the following, X is a technology that allows to conditionally forget or being absent-minded (which can be interpreted as another type of forgetting) while maintaining the requirement that $>_i$ is transitive. X can be customized to choose those actions that the player wants to forget and those that he wants to remember. We assume that X can only be used once.

Formally, if player i_0 has access to X for game Γ with perfect recall this transforms the game into a new game Γ^X where after each history, player i_0 may choose the action to take X, and taking X has an effect on the information sets that i_0 subsequently traverses. The set of histories is

$$\mathcal{H}^X = \mathcal{H} \cup \{(h, X) : h \in \mathcal{H} \setminus Z\} \cup \{(a_1, \ldots, a_K, X, a_{K+1}, \ldots, a_L) : (a_1, \ldots, a_L) \in \mathcal{H}, K < L\},$$

utility does not depend on taking X, i.e. $u_i^X(h) = u_i^X(h \setminus X)$, and the information set partition $\mathcal{I}_{i_0}^X$ for player i_0 in Γ^X has the following properties:

---

[1] Subsequently, we follow the definitions of Osborne and Rubinstein (1994).

[2] We subsequently say that history $a_1, \ldots, a_K$ is an **extension** of history $a_1, \ldots, a_L$

[3] Player i remembers what he previously knew if for information sets $I, J \in \mathcal{I}_i$ with $I \succ J$ we have $\{h \in T : h \succ I\} \subset \{h \in T : h \succ J\}$, i.e. if a history is already excluded at a past information set then so it is at a subsequent information set. Player i remembers his past actions if we cannot further partition $\mathcal{I}$ by only relying on his own past choices.

[4] In-transitivity is not far-fetched as it seems: Sending transactions to a blockchain mempool is an example of a situation with ambiguity about future, past and present, as we do not know how nature (searchers, builders and proposer) orders transactions.

1. Histories in which i_0 doesn't take X are partitioned as before and distinguishable from histories in which he does take X, i.e. $\mathcal{I}_{i_0} \subset \mathcal{I}_{i_0}^X$, (not taking X doesn't modify information sets and the player remembers whether he took X or not),
2. If $h, h' \in \mathcal{H}_{i_0}^X$ are in different information sets, then $h \setminus X, h' \setminus X \in \mathcal{H}_{i_0}$ are in different information sets in the original game (if i_0 remembers something after taking X he also remembers it in the original game, taking X leads to a coarsening of information sets),
3. The ordering $>_{i_0}^X$ on information sets is transitive.
4. For histories $h, h' \in \mathcal{H}_{i_0}^X$ that are extended in the same way $(h, a_1, \ldots, a_K), (h', a_1, \ldots, a_K) \in \mathcal{H}_{i_0}^X$, if h and h' are in the same information set and $A(h, a_1, \ldots, a_K) = A(h', a_1, \ldots, a_K)$ then $(h, a_1, \ldots, a_K)$ and $(h', a_1, \ldots, a_K)$ are in the same information set (if i_0 forgets something he cannot remember it later)
5. any histories $h, h' \in \mathcal{H}_0^X$ in which i_0 takes X and that only differ in when X is taken, $h \setminus X = h' \setminus X$, are in the same information set (the effect of X is independent of when it is taken),

Consider Figure 1 for an example of a game and the corresponding game with X. Note that X allows to forget actions after having observed them, as well as committing to forget future observations. The fourth item of our definition is for the sake of conceptual simplification: it is satisfied in all of our subsequent examples, but it restricts the powers of X. It would be an interesting extension for future work to extend the power of X by allowing to make the functionality of X be a function of history, or even make several versions of X available that allows to customize its effect during the game.

3 Applications

Subsequently we go through a variety of examples to demonstrate the power of X.

Arrow's information paradox

You are an ~~American~~ USA agent in the Soviet Union wishing to defect. You hold the list of all US agent names and would like to trade that list for riches in Moscow. However, knowing the KGB would not hesitate to kill you to prevent their information being devalued after you tell them, you have to calculate: what information to tell them, do you lie or not, or do you pretend to appear uninformed to back out of the trade.

This situation is an instance of Arrow's information paradox Arrow (1972): the potential buyer of a piece of information wants to inspect its content before deciding if they should buy it. However, once the buyer has inspected the information, they already own it and cannot ”un-know” the information due to their inability to forget. The seller, knowing all of this, would then choose to either 1) hide the information or 2) use an IP lawyer (or some trusted third party) to intermediate the transaction.

If they choose to hide the information, then adverse selection will take effect: now buyers are paying the average price for information, and sellers of high-value information will leave the market because they will not get a good deal. As a result, buyers will have lower willingness to pay because

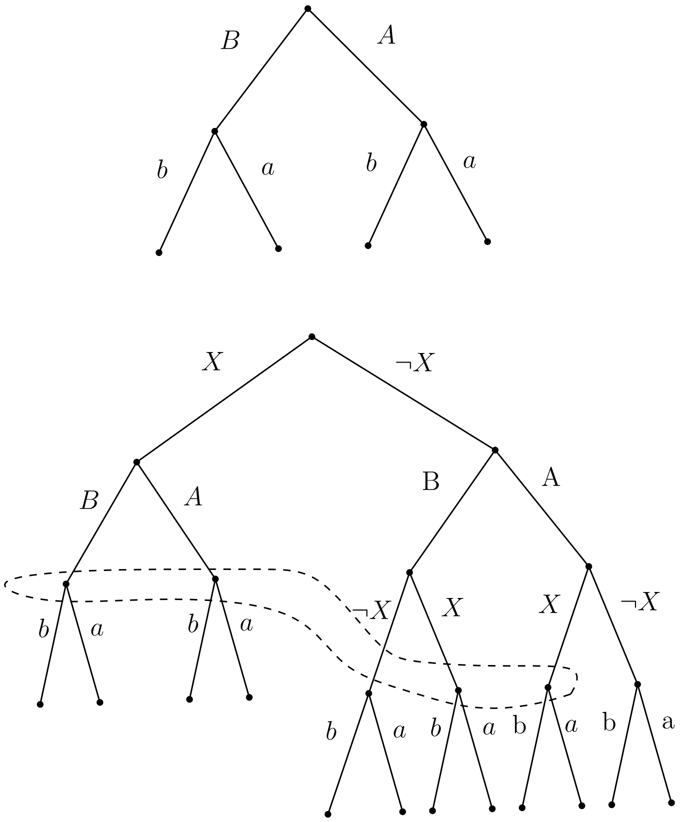


Figure 1: The original game (top): player 1 chooses between actions A and B and player between actions a and b. In the modified game where the second player has access to X (bottom), after each history, the second player can choose to take X which makes him forget whether the first action taken is A or B.

the average quality of the information has decreased. This "market for lemons" situation will persist until the market reaches a point where there is only valueless information and inevitably collapse.

Using lawyers and IP or patent law, on the other hand, creates high transaction costs, and a lot of the value in the information trade will be extracted by the lawyer and the legal system (assuming lawyers are scarce and they are the only source of credibility and mediation you could use for information transactions). Using lawyers requires, moreover, a functioning legal system to rely on, which might be a bit far-fetched in the spy example.

By administering pill X, the KGB can temporarily access information, confirm its veracity, and only remember the list after you have fled the country with their riches. Moreover, the KGB can mandate you to also ingest pill X, ensuring the information is purged from your memory. This mechanism mitigates the risk of subsequent unauthorized disclosures, which would otherwise devalue the information for the buyers. Consequently, pill X serves as a novel commitment device, potentially increasing the efficiency and reliability of high-stakes information markets.

Forgetting as substitute for commitment

Forgetting can often be used as a substitute for commitment. There are many examples of situations where a~~n~~ player may change their mind about an action once they receive new information, even though they would be better off when able to commit to a course of actions (which actions to take conditional on the content of the information) prior to receiving the information.[5]

The forgetful monopolist[6]

You run the local pharmacy and sourced a large supply of X. Each day the same customer enters your store and appears interested in buying X. You are not sure about his willingness to pay. Let's assume he needs at most one X per day and he cannot store X appropriately so that the X he buys today, he cannot use tomorrow. Of course if you sold X successfully to him today, then you know a lower bound on his willingness to pay tomorrow and could price X more accurately to extract more value from him ~~it seems.~~ For simplicity let's assume that the interaction happens in two periods (e.g. the day after tomorrow the customer will have forgotten about your pharmacy, this happens a lot since the invention of X). Suppose his willingness to pay is uniformly distributed on the unit interval and the same today as it is tomorrow.

In a Perfect Bayesian Equilibrium of the game you sell X for 3/10 today. If the customer buys X at the price of 3/10, you double the price tomorrow. If the customer doesn't buy X today, you leave the price at 3/10. The customer will buy X today if his willingness to pay is at least 6/10. Obviously he will buy X tomorrow if his willingeness is at least the offered price (6/10 or 3/10 depending on the situation). You will believe that the customer's ~~wtp~~ willingness to pay is uniformly distributed between 6/10 and 1 in case you sell X today and uniformly distributed between 0 and 6/10 in case you don't sell X today. It is straightforward to verify that this is a ~~PBE.~~ Perfect Bayesian Equilibrium

As you have a large supply of X you are of course wondering, would it be helpful for you to take X to forget that the customer who shows up today was already there yesterday. Why would that be? If you forget about the identity ~~about~~ of the customer you treat him each day as the typical customer with value uniformly distributed on the unit interval about whom you know nothing else. In that case you would sell X at the same price 0.5 each day. If you compare your revenue you will notice that you make $0.5 = 1/2 * 0.5 + 1/2 * 2 * 0$ on average if you are forgetful but only $0.45 = 3/10 * (3/10 + 6/10) + 4/10 * (0 + 3/10) + 3/10 * 0$ if you remember.

You are betrayed by your future self: it is in your best interest to extract a lot of rents once you have learnt that the customer has high willingness to pay. Thus, the customer is more careful in the present to hide his willingness to pay, even though it leads him to forgo an offer that he would otherwise accept. Your greedy future self hurts your bottom line today.

You should take X to make more money from selling X.

Reputation and Amnesia

Another class of examples where X can benefit the player is when it helps him ignoring the other player's attempt to coerce him into taking certain actions by building a (threatening) reputation.

---

[5] This has been recognized for example in the literature on persuasion.

[6] This is Wei Dai's story: http://www.weidai.com/monopoly-memory.txt

Bargaining and Amnesia

You go to the bazaar where the local carpet vendor offers you one of his magnificent rugs. The rug is advertised at a price of 10000\$. You are interested in the rug, you would in principle be willing to pay up to 12000\$ for it, but the offered price seems still excessive to you and you are convinced that the seller values it at 2000\$. However, you don't know whether the local carpet vendor is of the stubborn type who will stuck to the advertised price regardless of what you do (you believe that he is stubborn with prob π) or of the rational type who might settle for a smaller offer after negotiation. The uncertainty is mutual. The carpet vendor does not know whether you are the stubborn type who will stick to an initial offer of 4000\$ (he believes that you're stubborn with prob π) or whether you might settle for a higher price after negotiation. After you both have made y our initial offer, ~~we~~ you either have a deal (in case of mutually acceptable offers) or ~~we~~ you wait until one ~~of you~~ concedes and accepts the other's offer. This is the bargaining model with reputation of Abreu and Gul (2000). In the unique sequential equilibrium of the game you both choose a probabilistic strategy: You both initially mimic the stubborn type and offer to buy at 4000\$ resp. to sell at 10000\$. Afterwards you (and symmetrically the vendor) randomly concede where the probability of conceding before time $0 < t \leq T := -3\log(\pi)/\beta$ is

?

$$F(t) = 1 - e^{-\beta t/3}$$

where $0 < \beta < 1$ is the discount factor. A non-stubborn player will concede for sure by T since he is now convinced that the other player is stubborn.

The equilibrium above a) is inefficient (because it takes time to come to an agreement) and b) relies on reputation. Reputation assumes memory, put differently, reputation doesn't work on the forgetful. Can forgetting therefore restore efficiency?

Suppose, you have access to X. If you take the pill you remember the structure of the game, but you forget what has happened so far. Thus, you don't know for how long you have already been at the bazaar. You only observe that the game hasn't ended yet. Note that if the vendor would employ the equilibrium strategy above even if you have access to X, your beliefs would look very different: If you take X and forget, your belief will be stationary and therefore your action will be stationary.

In the situation where you have access to X but the seller doesn't we can construct the following equilibrium: You take X so that you continuously forget what happens during the game, you always offer to buy at 4000\$. The seller concedes independently of what you do. If he observes you taking X he concludes that you don't concede because you choose to not remember what he does. When the seller observes you not-conceding and not taking the pill, he concludes that you're of the stubborn type and also concedes.

Mafia and Collective Amnesia

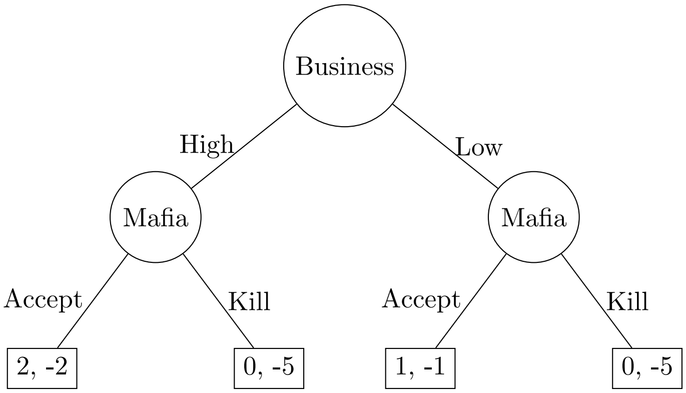
Sometimes forgetting doesn't help individually but can help collectively.

The local mafia is engaged in extortion activity. Each day they extort a different business. The scheme is as follows: the extorted business can propose a high or low extortion amount. The mafia can accept the offer or kill the business owner. There is uncertainty about the type of the mafia. The business owner fears that the mafia could be of the reckless type which always kills the business owner when he offers only a small amount (and accepts if the amount is high). The business owner assigns probability p to the mafia being reckless and probability $1-p$ of not being reckless. The non-reckless type of mafia behaves rationally and maximizes their total payoff. The business owner observes what the mafia has done in the past, murders are reported in the local newspaper in grueling details and it is easy to figure out (although not provable in a court) that the mafia is behind it. It is also easy to figure out whether any business has paid a large sum to the mafia, gossip tends to spread fast in the business community.

This is again a reputation game. It makes sense for the mafia to act recklessly to build a reputation of recklessness. More specifically, it will kill the business owner whenever it observes a low offer and accept when it observes a high offer. The business owner will make a high offer. It is important in this logic that there is a positive probability of the mafia being of the reckless type. This implies that if the mafia wouldn't kill the business owner for making a low offer, future business owners would learn that the mafia is of the rational type and would make a low offer.

More precisely, in the unique subgame perfect equilibrium, the (rational type) mafia plays "accept if high and kill if low" if so far no low offer has been accepted and "always accept" if in some previous period a low offer has been accepted. Each business plays "high" if so far no acceptance of a low offer has been observed and "low" otherwise. On the equilibrium path, business owners make high offers which are accepted by the mafia.

Now imagine the business owners have access to X. Individually this does not make a difference. But collectively, the business owners can profit from amnesia. Imagine the local business owners hold a meeting to which they invite the mafia to attend. At this meeting they collectively take X and this is observed by the mafia. After collective amnesia among business owners is established the game has a different subgame perfect equilibrium. In the equilibrium, the mafia "always accepts" and each business owner makes a low offer (now the business owners are not able to condition on the past).



Forward Induction and Amnesia

The pill also helps to eliminate forward induction reasoning in games, again through the channel of amnesia. You are the CEO of a company and wonder whether your company should enter a market with one incumbent. After careful market analysis you come to the conclusion that there are two possible strategies that you and the incumbent could follow. After entry, the company can focus on a niche inside that market or try to capture a large chunk of the market. Likewise, the incumbent can choose, after observing market entry by the new competitor, to serve the niche or the main segment of the market. If both the entrant and incumbent choose the same strategy they will both run a deficit. If they choose different strategies, serving the main segment of the market is more lucrative than choosing the niche. In particular, entering the market (which is costly) is only worthwhile for the entrant, if the incumbent serves the niche market and the entrant gets the main market after entry. There are two pure strategy subgame perfect equilibria of this game: in the first you choose to enter the market and capture the main segment of the market after entry. The incumbent, after observing entry will go into the niche. In the second, the CEO chooses not to enter the market, because he expects that conditional on entry the incumbent will capture the main segment of the market and only leaves the niche for his company to capture.

The second equilibrium is often ruled out by forward induction reasoning: if you choose to enter the market you signal that you expect that the incumbent goes into the niche, otherwise you wouldn't enter in the first place. Thus, the incumbent infers that the entrant whenever he enters will go for the main market and hence best replies by choosing the niche. In other words, the beliefs of the incumbent's CEO in the second equilibrium are not really compatible with his observations. He acts as if he would forget... This seems implausible.

Now suppose the incumbent's CEO has access to X. Forgetting now is very plausible. More precisely, we modify the game as follows: Nature randomly chooses at the beginning whether the game is a game of entry as described above or the game is one where the entrant is already in the market (the entry phase is not part of the game). Without the pill, the incumbent knows at which information set he is when making a decision and therefore knows whether they are in the entry game or in the game where the competitor has been in the market forever. Thus, in the subgame with entry the forward induction reasoning applies. If the incumbent has access to the pill and nature chooses the game without entry with sufficiently high probability, then forward induction reasoning has no power. We have an equilibrium where the incumbent gets the main segment and the entrant the niche and the entrant chooses not to enter the market.

One time use of information

Another powerful application of X is the enforcement of one time (or limited number of times) use of information. This is useful in variety of situations where sharing information can lead to a better outcome in a specific context, but agents might fear that the information is used in later situations to their detriment.

Information sharing in Bargaining

Bilateral bargaining is a classical situation where sharing information can increase efficiency because it makes an agreement more likely. However, the parties might not want the information to be used outside of the negotiation process. Thus, without the ability to conditionally share information, they might be less willing to share information during the negotiation process leading to a higher likelihood of bargaining failure.

Corporate

In corporate negotiations, non–disclosure agreements (NDAs) are signed to prevent leak of information and dictates that the information should not be used in a way that impacts the informants' behavior outside of the negotiation, it's essentially conditional recall but works at the natural language level. With pill X, there would not be a need for NDAs as negotiators would only be able to temporarily access private information and use it in the boardroom.

War

War should theoretically not occur in equilibrium unless there is incomplete information or a misalignment of preferences. In a scenario with complete information and alignment between decision-makers and those executing the war, countries would logically exchange resources rather than engage in conflict, as the outcome of the war would be predictable in advance. However, in reality, predicting the outcome of a war is challenging without full knowledge of the opposing side's military strength. Revealing such information—such as troop numbers or advanced weaponry—could compromise strategic advantages, making it unwise to disclose these details. This secrecy often results in mismatched expectations, leading both parties to a lose-lose situation by choosing war over negotiation.

In a hypothetical world with complete information, war might be avoided entirely. Jorge Luis Borges provides an apocryphal example[7]: ”In 15th century Italy, war had reached a perfection that many would call ridiculous. Once the armies were assembled, the generals compared the numbers, strength, and positioning of their troops and decided who among them must suffer defeat. Chance and bloodshed were eliminated.” So essentially, generals can mimic a real war by agreeing to all present a sample of their military at the same place at the same time, and then they can negotiate based on the partially-revealed information. A historical variant of this concept occurred in the Battle of the 300 Champions, where two Greek states each sent 300 elite soldiers to fight on behalf of their armies, avoiding full-scale conflict.

With pill X and conditional recall, it's possible to implement a mechanism where each country send their general to the other country's military base, review all of the firearms, and then negotiate the result of the war. This prevents either general from launching a followup attack after learning the other general's tactics and using that knowledge as an unfair advantage. Without the ability to do conditional recall, generals would not choose to reveal secret weapons or their surprise attack tactics because the other party may choose to dishonor the negotiation after learning those and think they can win the war with ~~those~~ that knowledge. This would then lead to both generals withholding information and therefore make the negotiation not honored and therefore ending in a total war. In Jorge Luis Borges's example, however, negotiation without pill X works perfectly because during Renaissance times the tactics are relatively straight forward, but the same setup wouldn't work in other conflicts.

Account delegation

You're planning a vacation and want to disconnect completely from the stress and noise of the internet. Knowing you wouldn't be able to resist the temptation of using your devices, you decide to leave them at your workplace. But a thought nags at you: What if something happens to my parents, and the hospital needs to call me? You try to convince yourself to bring the devices along, using this as justification.

---

[7] ”Libros y autores extranjeros”, book review of Erich Ludendorff's ”Der totale Krieg.”

Your friend, concerned about you sabotaging your trip, offers a solution: they'll monitor your calls, social media, and emails while you're away. If anything urgent arises, they'll handle it for you. Still, the idea makes you uneasy. What if they see messages I'm embarrassed about? Despite your friend's offer, your reluctance and anxiety lead to a predictable outcome—you spend the entire vacation doom-scrolling through Twitter.

Now, imagine your friend had access to X. The pill would let them filter and respond to only critical, family-related messages without exposing your private conversations. Similarly, you could think of this as using an LLM-powered email filter: raw data is processed, sensitive information is cleansed, and only essential details are passed along, all under a limited-use token. These types of applications are examples of ”one-time programs”Goldwasser et al. (2008).

Matching

In matching markets, it's tempting to misrepresent your traits to improve your matches. Over time, if everyone starts exaggerating or lying about their qualities, the market risks collapsing under the same principles seen in the ”Market for Lemons” example. Take a dating app, for instance: it's easy to claim you have desirable traits like a love for reading, a passion for traveling, or a good sense of humor. Verifying such claims, however, is costly and invasive. For example, proving someone's love of books might require examining their daily schedule or purchase history, while verifying shared tastes in movies could involve revealing a Netflix recommendation page. Understandably, most people are uncomfortable exposing such personal data without a certain level of commitment, leaving these traits unverifiable and the market susceptible to adverse selection.

X offers a potential solution. It allows users to safely share and verify sensitive information without overexposing themselves. For instance, a trusted system could review a person's reading habits or entertainment preferences, confirm their authenticity, and provide a simple ”stamp of approval” for specific traits. Adverse selection typically arises when verifying private information is too costly, invasive, or subjective. By enabling efficient and secure verification, X expands the range of traits that can be confidently validated, stabilizing the market and reducing the incentive to misrepresent oneself.

Spoilers

In the world of entertainment, spoilers can significantly impact one's enjoyment of a movie, book, or TV show. Currently, people rely on social norms and warnings to avoid spoilers, but these methods are imperfect. With pill X, one could safely engage in discussions about a piece of media without risking exposure to unwanted information. For instance, a fan could attend a QnA session with the creator of a TV series, gaining insights into the creative process without accidentally learning plot points they haven't reached yet. This would allow for more open and in-depth discussions within fan communities without the constant fear of spoilers, and we are already seeing real-life (https://chat.openai.com/g/g-Usa50OPO4-spoiler) implementations of this.

?

footnote

Acquisitions

In the business world, information about potential acquisitions can have significant unintended consequences. When a large company expresses interest in acquiring a smaller firm, the mere knowledge of this interest can cause the target company's stock price to spike, potentially making the acquisition more expensive or even unfeasible. "If oil company wants to buy your house, there is oil underneath." With pill X, negotiators could explore potential acquisitions without this information leaking to the market. They could conditionally recall the details of their discussions only when in specific negotiation settings, preventing premature market reactions and allowing for more efficient and discreet business dealings.

Juries & De-Biasing

You are called to jury duty in a murder trial. As part of the jury, your responsibility is to determine the facts of the case based on the evidence and testimony presented. However, not all evidence is admissible in court, even if it could potentially aid in uncovering the truth. For instance, evidence obtained illegally is excluded to ensure the integrity of the judicial process.

Now imagine a piece of inadmissible evidence comes to your attention. This information, while potentially helpful, might taint your judgment and conflict with your duty to base your decision solely on the admissible evidence. It's natural to want to uncover the unconditional truth,[8] yet your role requires you to focus on the most likely truth implied by the legally permissible evidence.

Here, X could help resolve this conflict. By taking it, you could focus only on the admissible evidence, filtering out the influence of irrelevant or prohibited information. There are other applications of de-biasing (or biasing): an LLM which has been trained on data that is biased (or was obtained illegally, e.g. in breach of intellectual property) or human agents with learned biases can un-learn their biases. This has various applications from fairer job recruitment, credit scoring, medical diagnosis systems to better recommendation systems. Of course, the same technology can also be used for malicious intentions by introducing biases through withholding relevant information.

4 Conclusion

We have explored the strategic advantages of credible, conditional forgetting across various contexts, illustrating how the hypothetical Pill X can enhance efficiency in many strategic interactions with potentially far-reaching implications. While the ability of credibly forgetting is just a thought experiment for human agents, it is a realistic possibility for artificial agents, as we sketched in the introduction. Even more so, several of the applications, such as the account delegation example are already implementable or implemented in reality through a combination of LLMs and account encumbrance by cryptographic means such as TEEs. Although human agents may never possess the ability to credibly and conditionally forget, they can still benefit from delegating decision-making to artificial agents equipped with this capability. As we've shown, this approach can serve as a cheap substitute to traditional commitment devices that rely on the legal system. More intriguingly, it also opens up entirely new possibilities that go beyond the scope of conventional mechanisms, offering exciting opportunities for innovation in mechanisms design.

---

[8] The evidence on whether juries are more likely to acquit if they have seen inadmissible evidence that points to the guilt of the defendant is mixed, but seems correlated with the severity of the prosecuted crime. In experiments, jurors faced with a hypothetical case, have a bias (for acquitting) if faced with exonerating evidence, but not if faced with other evidence.

References

Dilip Abreu and Faruk Gul. Bargaining and reputation. *Econometrica*, 68(1):85–117, 2000.

Kenneth Joseph Arrow. *Economic welfare and the allocation of resources for invention*. Springer, 1972.

Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum. One-time programs. In David Wagner, editor, *CRYPTO 2008*, pages 39–56, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

Harold William Kuhn. *Extensive Games and the Problem of Information.*, pages 193–216. Princeton University Press, Princeton, 1953.

Martin J Osborne and Ariel Rubinstein. *A course in game theory*. MIT Press, 1994.

Klaus Ritzberger. Recall in extensive form games. *International Journal of Game Theory*, 28: 69–87, 1999.